УДК 004.89:004.4+004.934

А.И. Парамонов
Донецкий национальный университет, Украина
Украина, 83001, г. Донецк, ул. Университетская, 24

О проблеме интерпретации тестовой информации

A.I. Paramonov
Donetsk National Universit, Ukraine
Ukraine, 83001, Donetsk, Universitetskaya Str., 24

The problem of Text Interpretation

А.І. Парамонов
Донецький національний університет, Україна
Україна, 83001, м. Донецьк, вул. Університетська, 24

Про проблему інтерпретації тексту

В статье рассмотрен подход к проблеме интерпретации текстовой информации. Предложен метод обработки текстовых фрагментов на основе нечеткой концептуальной модели представления знаний. Использование скрытых знаний (контекста), полученных путем вывода на основе опыта, позволяет расширить область интерпретации текста, а как следствие, повысить эффективность автоматизированных систем категоризации, аннотации, а также информационно-поисковых систем.
Ключевые слова: модель представления знаний, интерпретация текста, контекст.

In article the approach to a problem of text interpretation is considered. The method of text fragments processing on the basis of fuzzy conceptual model of knowledge representation is offered. Use of the tacit knowledge (context) received by a conclusion on the basis of experience, allows to expand area of text interpretation, and as a result, to increase efficiency of the automated systems of categorization, annotations and information search systems.
Key words: model of knowledge representation, the text interpretation, context.

У статті розглянуто підхід до проблеми інтерпретації текстової інформації. Запропоновано метод обробки текстових фрагментів на основі нечіткої концептуальної моделі представлення знань. Використання прихованих знань (контексту), отриманих шляхом виведення на основі досвіду, дозволяє розширити сферу інтерпретації тексту, а як наслідок, підвищити ефективність автоматизованих систем категоризації, анотації, а також інформаційно-пошукових систем.
Ключові слова: модель представлення знань, інтерпретація тексту, контекст.

Введение

Как самостоятельное научное направление искусственный интеллект (ИИ) существует сравнительно недолго, но уже достигнуты существенные результаты, сформированы некоторые концептуальные модели, устоялись определенные фундаментальные парадигмы. Исследования ИИ влились в общий поток технологий сингулярности, таких как информатика, нанотехнология, молекулярная биоэлектроника, квантовая теория и т.д. И уже не вызывают удивления идеи, что именно эти исследования будут определять характер того информационного общества, которое уже приходит на смену индустриальной цивилизации.

Одной из насущных и актуальных задач в современном информационном обществе является процесс накопления и применения корпоративных знаний. Информационные потоки, циркулирующие в информационно-аналитических центрах, преимущественно представляют собой неструктурированную разноязычную текстовую ин-

формацию. Принципиальной особенностью задач анализа текстовой информации является то, что предметом анализа выступают знания о предметной области, содержащиеся в текстовой информации. Таким образом, ядром системы поддержки информационно-аналитической деятельности должна быть система автоматизации распознавания, извлечения и формализации знаний, содержащихся в текстах, т.е. система понимания и интерпретации текстовой информации.

Решение поставленной задачи лежит на стыке ИИ и когнитивной психологии. Психолингвистические и психосемантические теории понимания текста предполагают, что в процессе понимания формируется смысл текста, отличный от исходного.

Для создания систем интерпретации текстовой информации предлагается концептуальная модель представления знаний [1], которая сохраняет содержащиеся в тексте и генерирует на их основе новые знания. Концептуальная гибридная модель представляет собой систему знаний, которая содержит множество базовых элементов (объекты, действия, события) и связи между ними (классификационные структуры). Предложенная система знаний позволяет преображать экстенсиональные представления, выраженные фрагментом текста, в интенсиональные представления системы, с возможностью появления новых представлений (извлечение из текста неявных знаний). В модели используется концепция прототипов, что позволяет учитывать особенность восприятия мира человеком в зависимости от его познаний и окружения. Для учета особенностей человеческого мышления и неоднозначности восприятия информации гибридная концептуальная модель формализована на основе аппарата нечетких множеств.

Компьютерные эксперименты, проведенные с предложенной моделью, подтвердили возможность представления знаний на ее основе и показали эффективность ее использования. В результате анализа экспериментов выявлена зависимость уровня понимания текста от предметной области. Так, для однозначной в плане терминологии области (например, «финансовые рынки») уровень извлекаемых знаний из корпусов текстов выше, чем для «многомерных» областей (например, «литературные произведения»). Это связано с наличием большого числа терминов с не-взаимно-однозначным сопоставлением формы и содержания, таких понятий, как синонимы, антонимы, омографы, полисемия, узуальные значения. Однако согласно опытам когнитивных психологов подобные различия в восприятии разной тематики наблюдаются и у основного числа испытуемых людей. Таким образом, для полноценного использования моделей представления и интерпретации текстовой информации, в том числе и разработанной нечеткой гибридной модели, необходим механизм однозначного сопоставления терминов с их смысловым содержанием, вкладываемым в них автором текста или воспринимаемым читателем.

Целью данной работы является разработка метода обработки текстовой информации на основе нечеткой концептуальной модели представления знаний.

В общем случае задача интерпретации текста на естественном языке сводится к преобразованию входного текста, который представлен в виде уровней конкретизации смысловой нагрузки, в элементы нечеткой концептуальной модели (рис. 1). Из известных моделей ИИ в основу интерпретации текста положен механизм рассуждения на основе опыта, что позволило использовать существующую систему знаний в условиях заданной предметной области, и благодаря этому определять адекватность и предметную направленность текста. Применение вывода на основе опыта позволило получить скрытые знания (неявные знания или смысл), содержащиеся в текст.

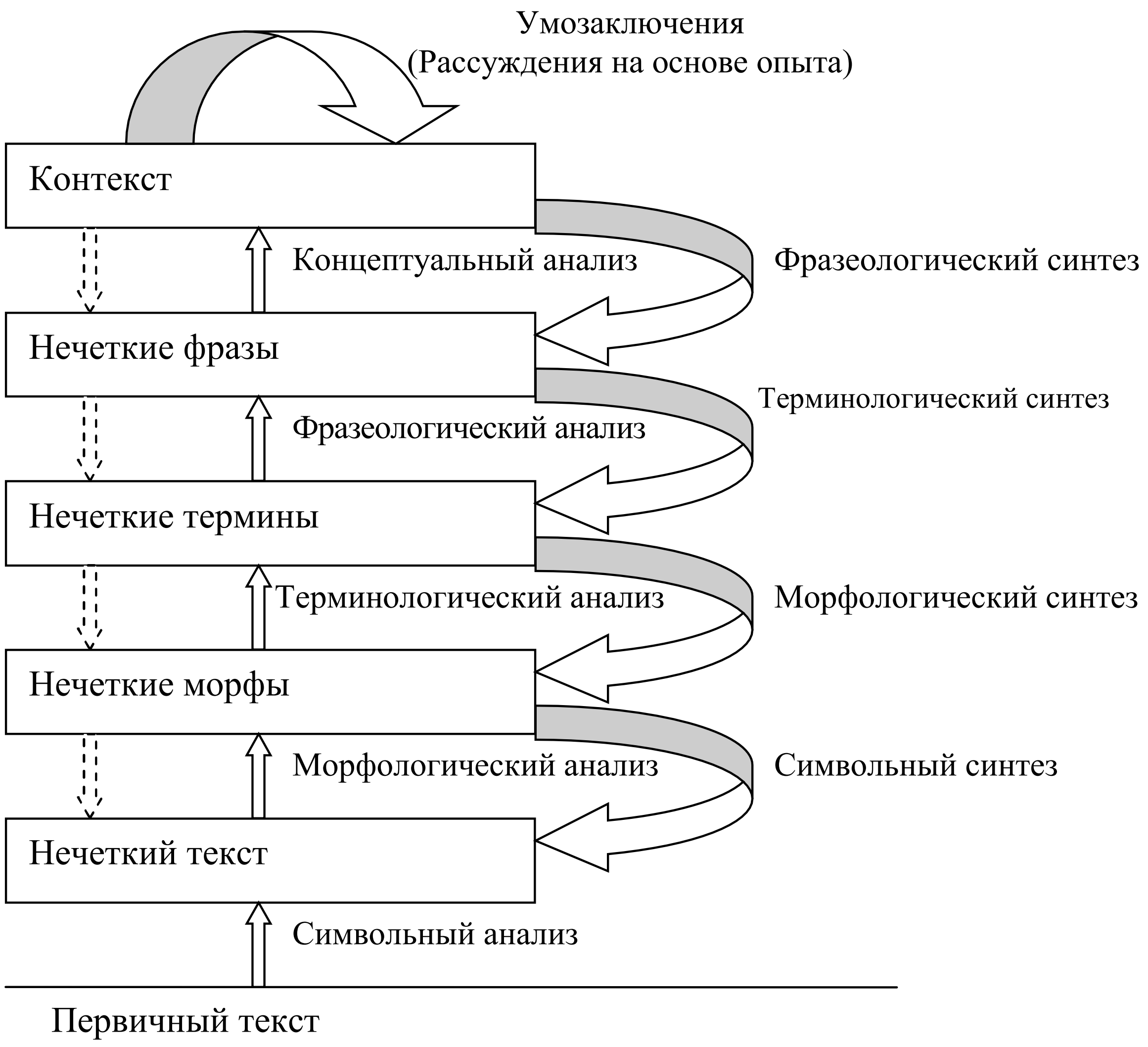


Рисунок 1 – Процесс обработки текстовой информации

Входная обрабатываемая конечная последовательность символов алфавита, которая может содержать ошибки, названа первичным текстом. Начальной обработкой входного текста является терминологическая разметка, в результате выполнения которой формируется подмножество терминов с некоторой степенью уверенности присутствия их в первичном тексте (нечеткие термины) [2]. Затем выполняется событийная разметка, которая предполагает построение последовательности фраз из нечетких терминов, собранных в определенные структуры (нечеткие фразы). Нечеткие фразы в свою очередь являются предметом концептуального анализа текста. Для интерпретации терминов (построение нечетких фраз) и отображения текста в виде событийной модели (построение контекста) используется предложенная концептуальная гибридная модель (ГМ) представления знаний.

Задачу однозначного сопоставления терминов с их смысловым содержанием предлагается решать на основе событийной разметки. На этапе событийной разметки нечеткие термины сопоставляются с элементами ГМ. Нечеткая характеристика термина задает фактор уверенности сопоставленного ему элемента. Одновременно с этим выполняется синтаксический анализ отдельного предложения, в которое входят рассматриваемые нечеткие термины. По сути, каждый рассматриваемый фрагмент первичного текста выступает отдельным предложением естественного языка. При сопоставлении термина с узлами классификационных сетей ГМ учитываются также и результаты грамматического анализа – определяется часть речи, которую он представляет. Соответственно, глаголы (сказуемые) активизируют узлы семантической сети действий, а

существительные (подлежащее, дополнение) активизирует узлы сети объектов. Сопутствующие части речи (например, прилагательные, обстоятельства) выступают в качестве свойств объектов (набор дополнительных характеристик). В качестве свойств узлов действий выступают предлоги, падежи и склонения глагола, которые указывают направленность воздействия, временные и другие характеристики.

Активность узла *i* (фактор уверенности в узле *i*) есть нечеткое множество $\underset{\sim}{X}_i$ (1), которое задается на базовом универсальном множестве X, введенном на шкале *[0; 1]*.

$$\underset{\sim}{X}_i = \{x|\mu_{\underset{\sim}{X}_i}(x)\},\ x \in [0;1] . \quad (1)$$

При активизации узлов значениям их активности сопоставляются значения уверенности в соответствующих терминах (2).

$$\mu_{\underset{\sim}{X}_i}(x) = \mu_{\underset{\sim}{Tj}}(x),\ x \in [0;1] . \quad (2)$$

После этого активизируются узлы-синонимы. Передача активности от узла к узлу выполняется по дуге, соединяющей эти узлы. Значение активности $[\underset{\sim}{Y}_{ji}]$, поступившее по дуге a_{ji} от узла *j*, рассчитывается по формуле (3).

$$[\underset{\sim}{Y}_{ji}] = [\underset{\sim}{X}_j] \times \varphi(a_{ji}), \quad (3)$$

где $[\underset{\sim}{X}_j]$ – активность смежного узла *j*, $\varphi(a_{ji})$ – поток по дуге a_{ji} [1], знак «$\times$» означает умножение нечеткого подмножества на число.

Активность передается только тем смежным узлам, в направлении которых дуги имеют значения потока больше заданного порога (экспериментальным путем определен порог «0,9»). То есть узлам, которые считаются синонимами. Текущая активность узла рассчитывается как общая накопленная активность $[\underset{\sim}{Y}_{ji}]_t$, поступившая от всех смежных узлов. В качестве механизма расчета текущей активности узла используется модель логогена Мортона [1], [3]. Начальное состояние активности узла задается нечетким множеством, характеризующим первоначальное отсутствие активности. Входом логогена является последовательность нечетких множеств, описывающих уровень активности смежных воздействующих узлов (3). Выходом является активность рассматриваемого узла после воздействия на него близких по смыслу узлов. В качестве операции изменения активности используется:

$$[\underset{\sim}{X}_i^l]_t = [\underset{\sim}{X}_i^{l-1}]_t * \lambda \times [\underset{\sim}{Y}_{li}]_t;\quad l = \overline{1,n}, \quad (4)$$

где $[\underset{\sim}{X}_i^{l-1}]_t$ – накопленная ранее активность узла (активность узла *i* до поступления активности от *l*-го узла), $[\underset{\sim}{Y}_{li}]_t$ – значение активности вновь поступившего свидетельства от *l*-го узла (3), λ – функция ограничения распространения активности (5), *n* – количество воздействующих узлов на узел *i*, ***** – операция накопления активности, реализованная в виде пересчета значений абсцисс и ординат *L-R* функции принадлежности для активированных узлов.

Таким образом, значения абсцисс *L-R* функции принадлежности для узлов, активированных при распространении активности, при перерасчете формул вычисляются следующим образом:

$$x_{[\underset{\sim}{X_i^l}]_t} = x_{[\underset{\sim}{X_i^{l-1}}]_t} + x_{[\underset{\sim}{X_l}]} \times (1 - x_{[\underset{\sim}{X_i^{l-1}}]_t}) \times \lambda(x) \; .$$

Соответственно, значения ординат рассчитываются в виде

$$\mu_{[\underset{\sim}{X_i^l}]_t}(x) = \mu_{[\underset{\sim}{X_i^{l-1}}]_t}(x) + \mu_{[\underset{\sim}{Y_{li}}]_t}(x) \times (1 - \mu_{[\underset{\sim}{X_i^{l-1}}]_t}(x)) \times \lambda(x);$$

$$\mu_{[\underset{\sim}{Y_{li}}]_t}(x) = \min\ (\varphi(a_{li}), \mu_{[\underset{\sim}{X_l}]}(x)) \; .$$

Введение функции λ позволяет контролировать процесс распространения активности.

$$\lambda(x) = \begin{cases} 1, & если \quad \varphi(a_{li}) \geq \delta \; ; \\ 0, & если \quad \varphi(a_{li}) < \delta \; , \end{cases} \tag{5}$$

где δ – пороговое значение распространения активности (для узлов-синонимов $\delta = 0{,}9$).

Когда все возможные трактовки терминов первичного текста определены, и все соответствующие узлы сетей объектов и действий активизированы, то активность передается на событийную модель – активизируются узлы пропозициональной сети. При этом семантическая сеть объектом будет иметь два активных подграфа, характеризующих объект и субъект действия.

Фактор уверенности в каждом из элементов события формирует корпусную модель фразы в виде возможных интерпретаций событий с заданной функцией уверенности в каждом событии. При этом множество активизированных событий зависит от модели мира в базе знаний (заложенных возможных конструкций событий) и от множества активизированных узлов семантических сетей. Таким образом, задается нечеткая фраза $\underset{\sim}{F_j}$ (6) как некоторое событие с заданной функцией уверенности в нем.

$$\underset{\sim}{F_j} = \{S_j, \underset{\sim}{\Theta_j}\}, \tag{6}$$

где S_j – *j*-е событие, $\underset{\sim}{\Theta_j}$ – фактор уверенности в *j*-м событии (7).

$$\underset{\sim}{\Theta_j} = (\underset{\sim}{\Theta_j^o} \cap \underset{\sim}{\Theta_j^c}) \cap \underset{\sim}{\Theta_j^д}, \tag{7}$$

где $\underset{\sim}{\Theta_j^o}$ – фактор уверенности в объекте *j*-го события, $\underset{\sim}{\Theta_j^c}$ – фактор уверенности в субъекте *j*-го события, $\underset{\sim}{\Theta_j^д}$ – фактор уверенности в действии (АКТе) *j*-го события. Факторы уверенности в каждом элементе события есть не что иное, как активность соответствующего узла классификационной структуры объектов или действий (3).

Из множества полученных нечетких фраз отбираем с наибольшим фактором уверенности (или с фактором уверенности, превышающим заданный порог). В результате фрагмент обрабатываемого текста в модели будет представлять собой последовательность входных событий (нечетких фраз) $\underset{\sim}{F_1}, \underset{\sim}{F_2}, \ldots, \underset{\sim}{F_k}$.

Фрагмент входной текстовой информации интерпретируется с учетом контекста. Контекст представляет собой активный подграф гибридной модели. Обработка оче-

редной нечеткой фразы фрагмента текста заключается в пересчете активности узлов подграфа гибридной модели (контекста). В основе подграфа положены узлы пропозициональной сети. Основной процедурой обработки знаний выступает распространение активности по сети. Активность узлов задаётся аналогично формуле (1).

В процессе интерпретации входных знаний у определенных узлов сети значения активности изменяются. При этом перерасчет происходит при обработке каждого события фрагмента текста. Значение активности узла при обработке *m*-го события фрагмента текста рассчитывается с учетом трех факторов:

1) *m*-го события фрагмента текста;
2) активности узла, полученной при обработке *m-1, m-2, …, m-k* событий фрагмента текста ($k<m$);
3) контекстных знаний.

Таким образом, значение нечеткого множества активности в момент времени *t* определяется как:

$$[\underset{\sim}{X}_i]_t = [\underset{\sim}{X1}_i]_t \oplus [\underset{\sim}{X2}_i]_t \oplus [\underset{\sim}{X3}_i]_t, \qquad (8)$$

где $[\underset{\sim}{X1}_i]_t$ – активность, поступившая при обработке *m*-го события, $[\underset{\sim}{X2}_i]_t$ – активность узла *i*, полученная при обработке предыдущих событий текста (составляющая активности от памяти), $[\underset{\sim}{X3}_i]_t$ – активность, поступившая от контекста (от смежных узлов сети), знак «$\oplus$» означает дизъюнктивную сумму нечетких подмножеств [4].

Каждая из трех составляющих оказывает воздействие на множество $\underset{\sim}{X}_i$ в один и тот же момент времени *t*. Процесс обработки входного фрагмента текста представляет собой дискретный во времени и непрерывный по состоянию процесс. Каждое изменение временного шага связано с обработкой очередного события фрагмента текста.

Активность, поступившая при обработке *m*-го события, есть отображение нечеткой характеристики входного события (нечеткой фразы) в активность узла.

С каждым узлом связана память глубиной *k*, где хранятся *k* значений активности, полученных на предыдущих *m-l* ($l = \overline{1,k}$, $k<m$) этапах обработки текста. Активностью узла *i* в контекстной памяти будем считать нечеткое множество, полученное как выпуклая комбинация нечетких множеств (активности на предыдущих этапах обработки) [4].

Под воздействием контекстных знаний понимается распространение активности внутри сети.

Распространение активности подразумевает, что на уровень активности узла воздействует активность, передаваемая ему от смежных с ним других узлов сети. Активность, поступившая от смежных узлов, рассчитывается по формуле (3).

При этом следует учесть, что в пропозициональной сети распространение активности возможно только по дугам, описывающим такие типы связей, как «во время» и «время» [1].

Отдельно следует упомянуть проблему «связности» текста, в основе которой – обработка цепочки событий по дугам «затем». При обработке текста (при формировании контекста) необходимо учитывать, как связаны во времени события, рассматриваемые на предыдущем этапе обработки и на текущем. Эта задача будет рассмотрена в следующих работах.

Составляющая активности узла от контекста рассчитывается как общая накопленная активность, поступившая от всех возможных смежных узлов. В качестве механизма расчета составляющей активности узла от контекста используется модель логогена Мортона, рассмотренная ранее. Рассчитывается по формуле (4).

После нескольких итераций волна распространения активности затухает. На этом этап воздействия *m*-го события фрагмента текста на контекст модели завершается. При обработке следующего события вычисления повторяются. Полученное подмножество активных узлов будет формировать контекстные знания модели о поступившем тексте.

Текущая активность узла определяет наличие элемента, который этот узел представляет, в обрабатываемом фрагменте текста.

На следующем шаге итерации первого такта текущая активность будет представлена как активность верхнего слоя памяти, и будет учтена при расчете новой активности.

«Смыслом» анализируемого текста (или контекстом) будем считать цепочку событий (фрагмент пропозициональной сети), узлы которой обладают нечеткой характеристикой активности.

В терминах ГМ все знания, содержащиеся во входной текстовой информации, будут представлены множеством *T*:

$$T = \{S_i, \tilde{X}_i\};$$

$$\forall \tilde{X}_i \in T \Rightarrow \exists(\mu_{\tilde{X}_i}(x) > 0),\ x \in \{b,c\},$$

где S_i – *i*-й узел пропозициональной сети, $\tilde{X}_i$ – активность *i*-го узла сети.

Для выделения контекста разного уровня восприятия текста введено понятие α -уровень активности, которое описывает обычное подмножество α -уровня нечеткого отношения [4]. В данном случае под нечетким отношением понимается активный подграф вида:

$$[\tilde{X}_i]^\alpha = \{x|\mu_{[\tilde{X}_i]^\alpha}(x)\};\quad \mu_{[\tilde{X}_i]^\alpha}(x) \geq \alpha;\quad \forall x \in \{b,c\},$$

где $[\tilde{X}_i]^\alpha$ – активный подграф α -уровня, α – пороговое значение активности узлов контекста ($\alpha \in [0;1]$).

Выводы

Представленная модель контекста может быть трактована как известный в литературе по искусственному интеллекту, но не формализованный метод рассуждений на основе опыта. Таким образом, рассуждения на основе опыта формализованы и представлены в виде модели изменения активности сети. Результаты вывода на основе неявных знаний (вывод на основе опыта), представленные в виде контекста, могут быть использованы для последующих рассуждений на основе поверхностных знаний (вывод на основе правил). Создание симбиоза в виде гибридной архитектуры, сочетающей в себе рассуждения на основе правил и опыта, позволяет расширить перечень задач, поддающихся автоматизации. Использование знаний, полученных путем вывода на основе опыта, позволяет расширить область интерпретации текста, а как следствие, повысить эффективность автоматизированных систем категоризации, аннотации, а также информационно-поисковых систем.

Литература

1. Парамонов А.И. Интенсиональные представления в виде нечеткой гибридной модели знаний / А.И. Парамонов // Искусственный интеллект. – 2008. – № 3. – С. 605-611.
2. Ломонос Я.Г. Терминологическая разметка текста в автоматизированной системе интеллектуальной обработки текстовой информации / Я.Г. Ломонос // Искусственный и интеллект. – 2006. – № 3. – С. 537-547.
3. Люггер Дж.Ф. Искусственный интеллект: стратегии и методы решения сложных проблем / Люггер Дж.Ф. – М. : Изд. дом «Вильямс», 2003. – 864 с.
4. Кофман А. Введение в теорию нечетких множеств / Кофман А. ; [пер. с франц.]. – М. : Радио и связь, 1982. – 432 с., ил.

Literatura

1. Paramonov A.I. Fuzzy hybrid model of knowledge as Intensional representation / A.I. Paramonov // Iskusstvennyj intellect. – 2008. – № 3. – P. 605-611.
2. Lomonos Y.G. Terminological text-markup into an automated system for intelligent processing of textual information / Y.G. Lomonos // Iskusstvennyj intellect. – 2006. – № 3. – P. 537-547.
3. Luger George F. Artificial intelligence: strategies and methods for solving complex problems / Luger George F. – Moscow : Publishing house «Williams», 2003. - 864 p.
4. Kofman A. Introduction to the theory of fuzzy sets [Translated from French.] / Kofman A. – Moscow : Radio and Communication, 1982. – 432 p., Ill.

RESUME

A.I. Paramonov

The Problem of Text Interpretation

In article the approach to a problem of text interpretation is considered. The general scheme of text processing is described. The method of interpretation textual information on the basis of fuzzy conceptual model of knowledge representation [1] is developed.

Determinations of fuzzy data representation are defined. The fuzzy characteristic of conceptual models networks – activity is entered. It is offered to use formalized cognitive Morton’s logogen model (4). The stage of the conceptual analysis is constructed on the basis of networks activity model. The model of activity assumes that the sense of the text is formed as set of three components (8): the processed text, memory and contextual knowledge. The mechanism of reasoning on the basis of experience is formalized in the form of text interpretation model.

Results of a conclusion on the basis of the tacit knowledge (a conclusion on the basis of experience), which are presented by a context, it’s offered to use for the subsequent reasoning on the basis of superficial knowledge. Use of contextual model allows to expand area of text interpretation, and as a result, to increase efficiency of the automated systems of categorization, annotations and information search systems.

Статья поступила в редакцию 05.06.2012.